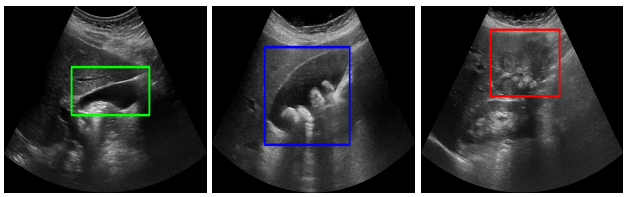
Surpassing the Human Accuracy: Detecting Gallbladder Cancer from USG Images with Curriculum Learning (Supplementary Material)

Soumen Basu[1], Mayank Gupta[1], Pratyaksha Rana[2], Pankaj Gupta[2], Chetan Arora[1]
[1] Indian Institute of Technology, Delhi, India
[2] Postgraduate Institute of Medical Education and Research, Chandigarh, India
https://gbc-iitd.github.io/gbcnet

A. Details of Data Acquisition and Annotation

Data Acquisition: The study was approved by the ethics committee of the Postgraduate Institute of Medical Education and Research, Chandigarh. We performed all procedures according to the Declaration of Helsinki and the research guidelines of Indian Council of Medical Research. According to the hospital's protocol, 6 hours fasting was advised a day before the Ultrasound (USG) examinations for adequate distension of the GB. Two radiologists with expertise in abdominal USG performed the examinations on a Logic S8 machine (GE Healthcare) using a convex low-frequency transducer with a frequency range of 1–5 MHz. USG assessment was done from different angles using both subcostal and intercostal views to visualize the entire GB, including the fundus, body, and neck. Patients were examined in different positions for adequate visualization of the GB. The screen area was adjusted so that the GB could occupy at least 20% of the entire screen.



(a) (b) (c)

Figure S1. Sample ROI annotation. (a) Normal GB with ROI annotated in green, (b) GB with benign abnormalities with ROI in blue, and (c) Malignant GB with ROI annotated in red.

ROI Annotation: Apart from image classification labels, we used bounding-box annotations to capture the GB localization. Two radiologists with 7 and 2 years of experience in abdomen radiology did the bounding-box annotations with consensus using the LabelMe [7] software. A single free-size axis-aligned rectangular box in every image, spanning the entire GB and adjacent liver parenchyma, preferably keeping the GB in the box's center, highlights the region of interest (see Fig. S1).

Model	Acc	Spec.	Sens.
ROI+VGG16	53.3 ± 9.2	71.9 ± 11.5	73.3 ± 17.9
ROI+VGG16+VA	77.7 ± 4.1	93.8 ± 3.0	72.0 ± 19.5
ROI+ResNet50	76.6 ± 10.7	82.3 ± 10.5	90.9 ± 11.1
ROI+ResNet50+VA	85.4 ± 7.7	92.3 ± 5.9	87.5 ± 9.1
ROI+Inception-V3	71.8 ± 8.9	83.3 ± 8.7	78.5 ± 21.4
ROI+Inception-V3+VA	82.6 ± 4.6	93.1 ± 4.4	82.6 ± 9.9
RetinaNet	74.9 ± 7.3	86.7 ± 7.8	79.1 ± 8.9
RetinaNet+VA	73.3 ± 6.0	92.1 ± 4.4	70.6 ± 14.2
GBCNet (ROI+MS-SoP)	88.2 ± 5.1	94.2 ± 3.7	92.3 ± 7.1
GBCNet+VA	92.1 ± 2.9	96.7 ± 2.3	91.9 ± 6.3

Table S1. Model performances (10-fold cross-validation) for training with our proposed visual acuity-based curriculum.

B. Performance Improvement with Proposed Curriculum

We show the performance improvement of various models with the curriculum-based training in Tab. S1. All models show improvement in specificity, which indicates the effectiveness of the proposed blurring-based curriculum in tackling texture bias.

C. Implementation details

Tab. S2 lists the configurations of all models which we have used. We trained on the Quadro P5000 16GB GPU. The table includes a brief description of the various stages of the network, input image sizes ($H \times W \times D$), the optimizer, relevant hyper-parameters such as learning rate, weight decay, momentum, batch size, and the number of training epochs/steps for the network.

Model	Description	Input Size	Optimizer	Batch size	Epochs/ Steps
YOLOv4 [1]	CSPDarknet53 backbone, PANet neck, anchor-based YOLO head. Total 162-layers. Backbone was frozen for first 800 step. Entire network was trainable thereafter. Single stage, anchor-based	$608 \times 608 \times 3$	SGD LR = 0.0001 momentum = 0.95 weight decay = 0.0005	64	3000 steps
Faster-RCNN [5]	Resnet50 Feature Pyramid backbone. Backbone was frozen for training. Two-stage, anchor-based.	$800 \times 1333 \times 3$	SGD LR = 0.005 momentum = 0.9 weight decay = 0.0005	16	60 epochs
Reppoints [11]	Resnet101 backbone, Group Normalization neck, and a reppoints head. Backbone was frozen for first 30 epochs, and entire network was trainable thereafter. Two-stage, anchor-free	$800 \times 1333 \times 3$	SGD LR = 0.001 momentum = 0.9 weight decay = 0.0001	4	50 epochs
Centripetal-Net [2]	Improvement over CornerNet model. Uses centripetal shift to match corners. HourglassNet-104 backbone. Enitre network was trainable. Anchor-free	$511 \times 511 \times 3$	Adam LR = 0.0005	4	50 epochs
ResNet [3]	Resnet-50 used. All layers were trainable. Output dimension of last fully connected layer is three - corresponding to normal, benign, and malignant GB. LR decays by 10% after every 5 epochs through a step LR scheduler.	$224 \times 224 \times 3$	SGD LR = 0.005 momentum = 0.9 weight decay = 0.0005	16	100 epochs
VGG [8]	VGG-16 is used. All layers were trainable. LR decays by 10% after every 5 epochs through a step LR scheduler.	$224 \times 224 \times 3$	SGD LR = 0.005 momentum = 0.9 weight decay = 0.0005	16	100 epochs
Inception [9]	Inception-V3 used. All layers were trainable. LR decays by 10% after every 5 epochs through a step LR scheduler.	$299 \times 299 \times 3$	SGD LR = 0.005 momentum = 0.9 weight decay = 0.0005	16	100 epochs
RetinaNet [4]	Resnet-18-FPN used as backbone. All layers were trainable. Three output classes corresponding to normal, benign, and malignant GB.	$512 \times 512 \times 3$	Adam LR = 0.0001	8	50 epochs
EfficientDet [10]	EfficientNet-B4 used as backbone and BiFPN as feature network. All layers were trainable. Three output classes corresponding to normal, benign, and malignant GB.	$1024 \times 1024 \times 3$	Adam LR = 0.001	2	50 epochs
MS-SoP Classifier (Ours)	16 MS-SoP layers. All layers were trainable. Three output classes corresponding to normal, benign, and malignant GB.	$224 \times 224 \times 3$	SGD LR = 0.005 momentum = 0.9 weight decay = 0.0005	16	100 epochs

Table S2. Implementation details for the different baseline networks used for classification and gallbladder localization.

D. Calculating Precision and Recall for GB Localization Networks

For computing precision and recall during the GB localization phase, as suggested by [6], if the center of the predicted region lies within the bounding box of the ground truth region, then we consider a region prediction to be a true positive; otherwise, we consider the region prediction to be a false positive due to localization error. Further, we consider the zero/no region prediction as a false negative (all our images contain GB, and the localization network's task is to merely localize it).

E. GradCAM Visuals for GBCNet

Figure S2 shows the sample Grad-CAM visualizations of the predictions using GBCNet (ROI+MS-SoP) with curriculum learning.

F. ROI Visuals

In figure S3, we show sample predictions of the GB region localization for different models. We also show the region of interest as perceived by the expert radiologists. The localization model is fairly accurate in capturing important regions of the USG image.

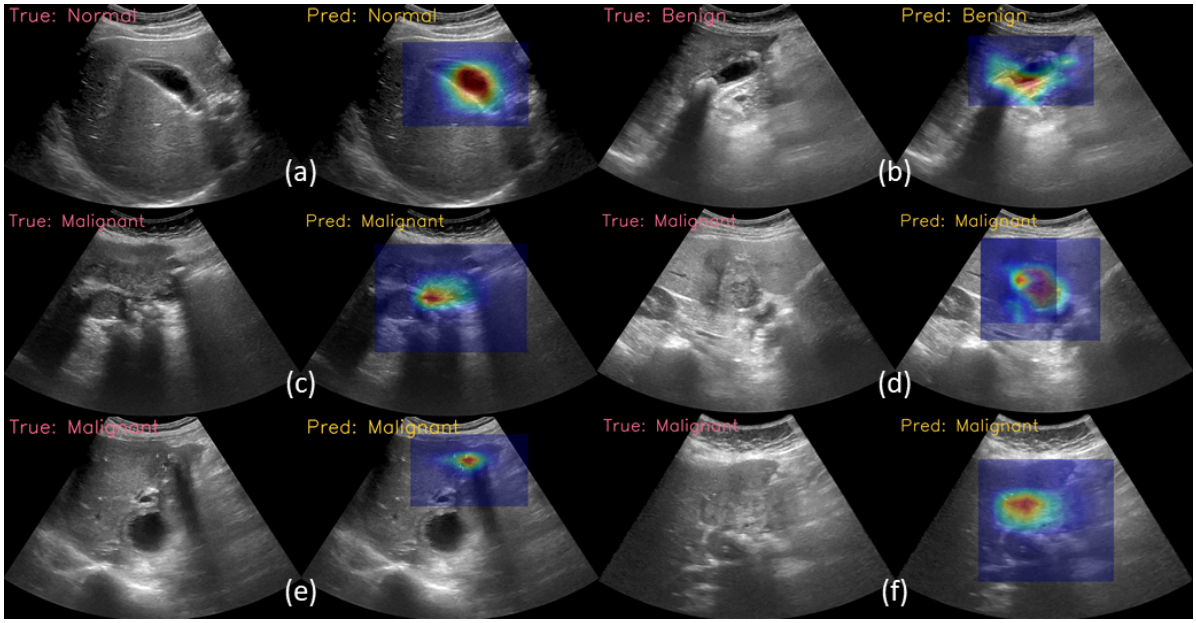


Figure S2. Sample Grad-CAM visuals of GBCNet with curriculum learning. (a) Normal, (b) Benign, and (c)–(f) Malignant samples.

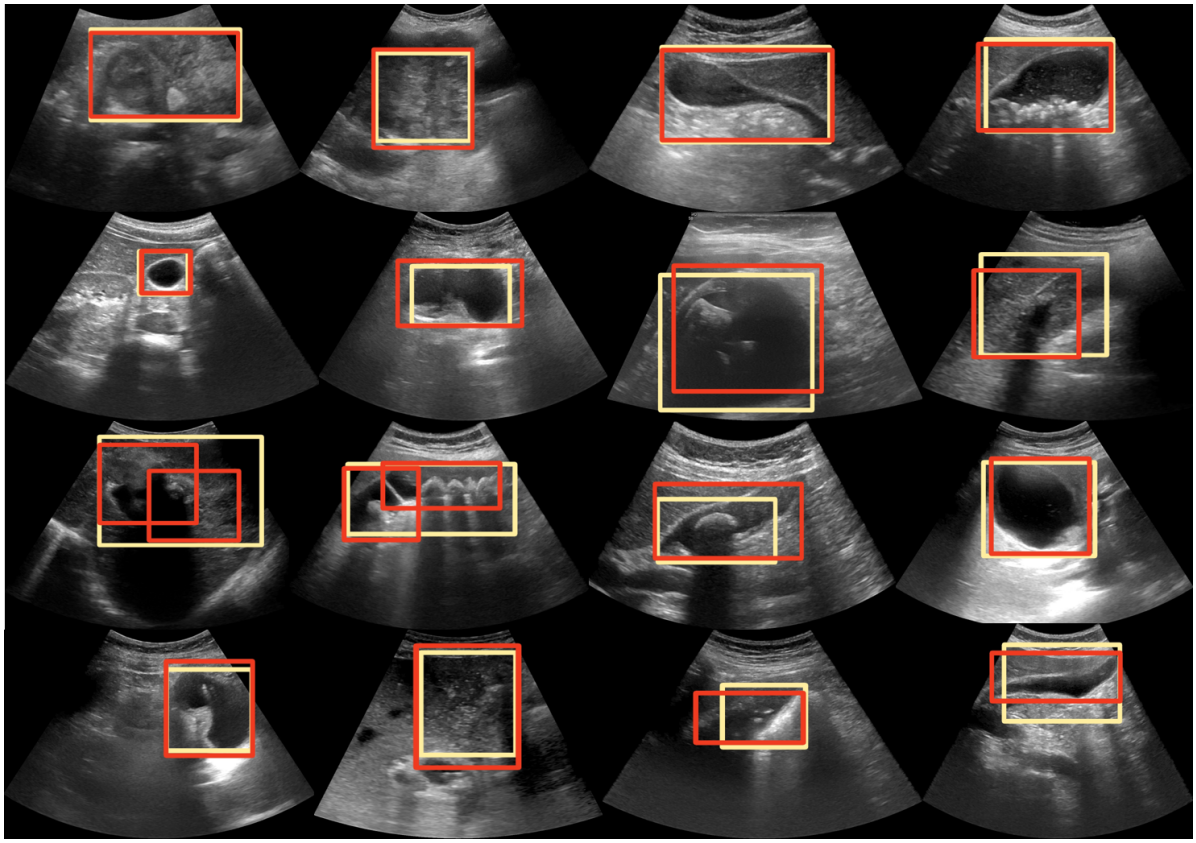


Figure S3. Sample visual results of RoI Detection models. First row - Faster-CNN, second row - YOLOv4, third row - Reppoints, and fourth row - CentripetalNet. Dark red is the ROI prediction by the model and light yellow is expert radiologists' perception of ROI.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[2] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. IEEE ICCV*, pages 2980–2988, 2017. 2

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 2

[6] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):1–7, 2018. 2

[7] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2

[10] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proc. IEEE CVPR*, pages 10781–10790, 2020. 2

[11] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 2